Automated Extraction of Biotransformation Data from Metabolic Schemas Using Geometric Criteria and Large Language Models

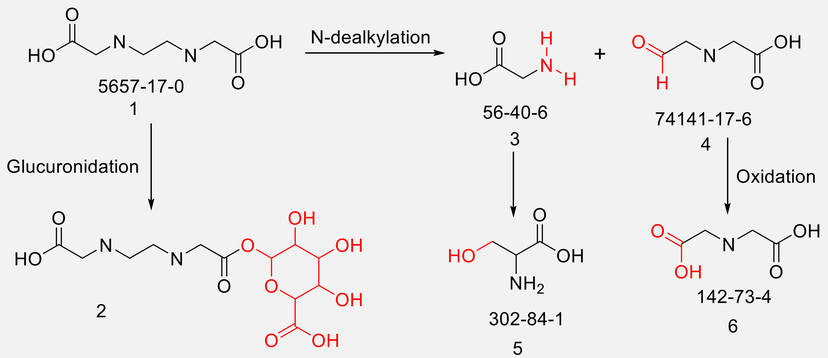
4067/K566

Oleg Stroganov[1], Jesse Gordon[1], David Merberg[1], ElLantae' Byrd[2,*], Ashley Mudd[2], Bastian Selman[2], Guilherme Soares[2], Gang Yan[2], Cathy Lester[2]
[1]Rancho BioSciences, PO Box 7208, Rancho Santa Fe, CA 92067; [2]The Procter and Gamble Company Mason, OH USA

Background

- Understanding metabolism is critical for safety assessment
- Metabolite prediction models often overpredict the number of metabolites
- We have summarized biotransformation pathways for > 2000 compounds
- This poster describes the digitization of metabolism data and population of the database housing the data
- Methods described here can be used to digitize chemical information and associated metadata

Metabolic Pathway



Metabolism report

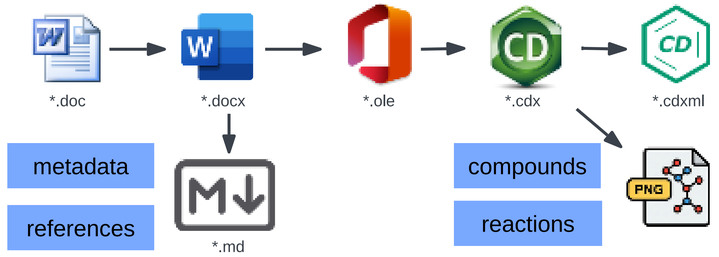
Compound (1) is Ethylenediamine-*N,N'*-diacetic acid. Compounds of this class (e.g. Edetic Acid) are poorly absorbed from gastrointestinal tract and are rapidly excreted unchanged in the urine.[1]

As compound (1) is a di-acid, it may undergo glucuronidation to form metabolites as shown in structure (2). Compound (1) may undergo oxidative *N*-dealkylation to form glycine (3) and metabolite (4). Glycine (3) is reported to be converted to serine (5) by Serine hydroxymethyltransferase (SHMT) or degraded to NH_3 and CO_2 by the Glycine cleavage enzyme system (GCS) in humans and animals.[2] Metabolite (4) can further undergo oxidation of the aldehyde functionality, resulting in formation of a diacid (6).

References:

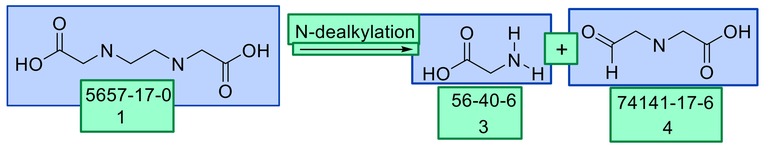
1. Edetic acid (EDTA) in Drinking-water. Background document for development of WHO Guidelines for Drinking-water Quality. *Guidelines for drinking-water quality*, 2nd ed. Addendum to Vol. 2. Health criteria and other supporting information. World Health Organization, Geneva, 1998.

Metabolism Data Extraction Pipeline



1. Metabolism reports are converted to docx and then to **markdown**. Chemical information is extracted with ChemDraw and saved to cdxml and png format
2. Metabolism **metadata** (biospecimen, species, reference type) and reference are extracted from markdown by LLM (claude-3.5)
3. Chemical **structures** and **reactions** are extracted from cdxml using custom parsing scripts that use geometric criteria to detect reaction direction. Compound numbers are taken from cdxml text fields.
4. Type of **biotransformation** is detected using SMIRKS patterns
5. Reactions are associated with metadata using compound numbers from schemas and text

Extraction of Chemical Information



1. Bounding boxes for compounds, text blocks and arrows are extracted from ChemDraw xml files
2. Compounds which are close to each other and have a "+" sign between them are joined in groups
3. Reactions are constructed using directions of arrows and position of tail and head relative to compounds
4. Reaction graph is analyzed, and inconsistencies (multiple roots, orphan compounds) are reported for manual review

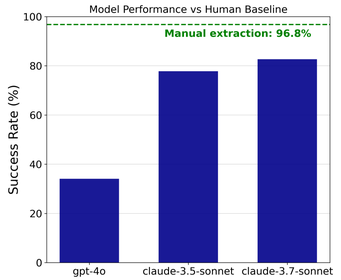
Extraction Quality Assessment

Quality of data extraction was assessed by comparing manually extracted data for a subset of reports with data extracted automatically.

Extraction task	Success rate	
	automatic	manual
Extraction of reaction graph	98.5%	96.8%
Extraction of reagents and products ids	98.9%	92.3%
Extraction of CAS numbers	93.4%	99.7%
Annotation of biotransformation type	79.8%	94.0%

Most frequent issues: incomplete annotation of biotransformations (automatic), incomplete extraction of compound numbers (manual), missing CAS numbers (automatic)

Biotransformation Schema Digitization with LLM



- PNG images from ChemDraw xml files were used to extract information with multi-modal LLMs.
- Anthropic models were more reliable in extraction of reaction graph, with success rate of 83%
- Typical issues were mistakes in processing reaction groups

Impact and Use

- Reduce the use of animals for generating metabolism information
- Make metabolism information available and AI-ready for predictive models
- Build metabolite prediction models addressing a wider chemical domain than currently available tools based on small pharmaceutical molecules

Rancho BioSciences

P&G Human Safety
Ensuring Safe Products